Pancaldi V, Carrillo-de-Santa-Pau E, Javierre BM, Juan D, Fraser P, Valencia A, Rico D.

Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity.

***Genome Biology* 2016, 17: 152**

DOI link to article:

https://dx.doi.org/10.1186/s13059-016-1003-3

Date deposited:

12/10/2016

Pancaldi et al. Genome Biology (2016) 17:152
DOI 10.1186/s13059-016-1003-3

Genome Biology

RESEARCH Open Access

Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity

Vera Pancaldi[1*], Enrique Carrillo-de-Santa-Pau[1], Biola Maria Javierre[2], David Juan[1], Peter Fraser[2], Mikhail Spivakov[2], Alfonso Valencia[1] and Daniel Rico[1*]

Abstract

Background: Network analysis is a powerful way of modeling chromatin interactions. Assortativity is a network property used in social sciences to identify factors affecting how people establish social ties. We propose a new approach, using chromatin assortativity, to integrate the epigenomic landscape of a specific cell type with its chromatin interaction network and thus investigate which proteins or chromatin marks mediate genomic contacts.

Results: We use high-resolution promoter capture Hi-C and Hi-Cap data as well as ChIA-PET data from mouse embryonic stem cells to investigate promoter-centered chromatin interaction networks and calculate the presence of specific epigenomic features in the chromatin fragments constituting the nodes of the network. We estimate the association of these features with the topology of four chromatin interaction networks and identify features localized in connected areas of the network. Polycomb group proteins and associated histone marks are the features with the highest chromatin assortativity in promoter-centered networks. We then ask which features distinguish contacts amongst promoters from contacts between promoters and other genomic elements. We observe higher chromatin assortativity of the actively elongating form of RNA polymerase 2 (RNAPII) compared with inactive forms only in interactions between promoters and other elements.

Conclusions: Contacts among promoters and between promoters and other elements have different characteristic epigenomic features. We identify a possible role for the elongating form of RNAPII in mediating interactions among promoters, enhancers, and transcribed gene bodies. Our approach facilitates the study of multiple genome-wide epigenomic profiles, considering network topology and allowing the comparison of chromatin interaction networks.

Keywords: Assortativity, 3D genome, Chromatin Interaction Network, Embryonic stem cells, Epigenomics, Promoter Capture Hi-C, Enhancers, Polycomb, RNA polymerase

Background

Advances in chromatin interaction mapping have allowed us to refine our vision of the genome, leading us to a more realistic, well organized tension globule picture with extrusions of chromatin loops [1, 2]. The resolution of available contact maps has increased from a megabase to less than a kilobase in just 5 years [3–10]. However, our understanding of what determines the three-dimensional (3D) structure and of its functional importance remains limited. Starting from the first papers modeling DNA as a polymer and the genome as a polymer globule [1, 2, 11], scientists have been looking for a connection between the chromatin contact configuration and the regulation of gene expression [12–14]. It is now accepted that gene regulation happens as much through distal enhancer elements as through proximal promoters and the distinction between promoters and enhancers has itself been put to the test [15, 16].

The combination of chromatin capture experiments with next-generation sequencing (NGS) has enabled the characterization of chromatin contacts at an unprecedented level of detail. Different techniques yield different

* Correspondence: vpancaldi@cnio.es; drico@cnio.es
Alfonso Valencia and Daniel Rico are co-senior authors.
[1]Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
Full list of author information is available at the end of the article

views of the genome. High-throughput conformation capture (HiC) is an unbiased approach that allows us to investigate the three-dimensional structure of the genome of given cell types [3, 9], even in single cells [17] during differentiation processes [10, 18–20] and across species [21, 22]. The HiC technique assays, in principle, all versus all chromosomal contacts, requiring very high sequencing coverage and making it very costly and practically almost impossible to achieve saturating coverage. Alternative approaches allow exploration of the contacts of a subset of genomic regions, with higher resolution at the same cost. For example, chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) [23] analyzes only those interactions that are mediated by a protein of interest by pulling down only the interacting fragments that include this protein.

Recently, other capture approaches were developed that enable selective enrichment for genome-wide interactions involving, at least on one end, specific regions of interest. For example, capture HiC was recently used to identify the chromatin interactions involving colorectal cancer risk loci [24]. A similar approach is used in promoter capture HiC (PCHi-C) [8], which detects both promoter–promoter interactions and interactions of promoters with any other non-promoter regions. These interactions are therefore identified irrespective of target promoter activity and across the whole range of linear genomic distances between fragments. HiCap [7] is a similar approach to detect promoter-centered chromatin interactions. The two methods provide a complementary view of chromatin interactions as PCHi-C yields larger fragments (average fragment size 5 kb) and longer interaction ranges (on average 250 kb), whereas HiCap has better resolution (average fragment size < 1 kb) but less coverage of long range interactions. Thanks to these new techniques, we can now use interactions between non-coding parts of the genome and genes to interpret the wealth of disease-associated genomic variation data which were so far unexplained [24–26].

The increasing availability of 3D interaction datasets for multiple cell types and organisms has prompted the development of multiple data processing approaches. Important factors need to be taken into account in these analyses: one is the detection of biologically significant interactions from the background noise of interactions purely due to the linear proximity of the two fragments on the genome; another is the averaging effect that is produced by the heterogeneity of contacts in different cells [27]. While various methods for normalizing and detecting signals in HiC-related datasets have been developed [28–30], downstream interpretation of the resulting contact maps represents a significant problem. Moreover, to this day, no single unified standards are available for these types of data, hindering the direct comparison between the chromatin structure in different cell types, species or conditions [28]. The field is moving fast, however, as shown by the recent focus on unraveling the 4D nucleome, that is, the internal organization of the nucleus in space and time, even at the resolution of single cells [31, 32].

Given the complexity of these datasets, it is intuitive and useful to represent them as networks in which each chromatin fragment is a node and each edge (link) represents a significant interaction between two chromatin fragments. This framework allows us to study the properties of the 3D chromatin structure using tools from network theory. The booming field of network science provides a useful toolbox and different metrics that can be used to compare and interpret chromatin contact networks from a more global point of view. For example, one can identify the most connected nodes or look for functional relationships between nodes that interact more than expected by chance [33].

A few previous papers have dealt with network analysis approaches applied to chromatin interaction networks [33–37], with the aim of unraveling general principles of 3D chromatin organization. For example, in the pioneering work by Sandhu et al. [35], the chromatin interaction network is constructed starting from RNA polymerase II (RNAPII) ChIA-PET performed in mouse embryonic stem cells (mESCs) to obtain a single large connected component. An accurate network analysis revealed the functional organization of different chromatin communities. A similar analysis, performed on the budding yeast chromatin interaction network, showed that cohesin mediates highly interconnected interchromosomal subnetworks (cliques) which are stable and have similar DNA replication timing [33].

In this work, we aim to establish which properties of the DNA or which factors bound to it can be associated with specific types of 3D chromatin contacts. To this end, we project the linear chromatin context information directly onto the 3D network, preserving its topology. We focus our analysis on mESCs as chromatin interactions for this cell type have been assayed by multiple techniques and a very comprehensive epigenetic characterization is available. We study interaction networks derived by state-of-the-art PCHi-C in mESCs, in which we quantify the assortativity of 78 chromatin features (three cytosine modifications, 13 histone modifications, and 62 chromatin-related protein binding peaks [38]).

In social sciences, assortativity is used to measure the extent to which similar people tend to connect with each other [39, 40]. Whereas in society it is easy to imagine which principles might lead people from the same ethnic origin or cultural background to establish social ties, we are still investigating principles that organize chromatin in the nucleus. We borrow the concept of assortativity,

making an analogy between social networks and chromatin contact networks, and introduce the concept of chromatin assortativity (ChAs). This global measure identifies to what extent a property of a chromatin fragment is shared by fragments that interact preferentially with it. If a feature appears to be localized in specific well-connected areas of the network, it will be characterized by having high ChAs. Identifying features with high ChAs can thus lead us to candidates for factors that might mediate chromatin contacts. This would be an important step forward in elucidating the organizing principles inside the nucleus and furthering our understanding of the mechanistic basis of genome regulation.

Polycomb group (PcG) proteins and associated marks have the highest ChAs values, imposing themselves as the factors that are more strongly related with chromatin structure in mESCs, as recently suggested [5, 20, 41]. Through this novel analysis, we also gain insight regarding different RNAPII variants as important players shaping the 3D chromatin structure. More specifically, we note a different configuration of actively elongating RNAPII forms in promoter–other end contacts compared with non-elongating RNAPII variants. This finding is confirmed in three independent datasets and it suggests that actively elongating RNAPII is involved in the contact between regulatory elements and their targets.

Results

The chromatin interaction network

To assemble the chromatin interaction network, we used the recent PCHi-C dataset in mESCs from Schoenfelder et al. [8], including interactions amongst promoters and between promoters and other genomic elements. The PCHi-C data were processed using the CHiCAGO algorithm. CHiCAGO is a HiC data processing method that filters out contacts that are expected by chance given the linear proximity of the interacting fragments on the genome and takes into account the biases introduced by the capture step used in the PCHi-C approach [29]. The network based on the significant interactions detected by CHiCAGO has 55,845 nodes and 69,987 connections (see "Methods" and Additional file 1: Figure S1). Of these interactions, 20,523 interactions connect a promoter fragment with another promoter fragment (P–P edges) and 49,464 interactions connect promoters with non-promoter "other end" fragments (P–O edges).

As in many networks, we can observe a main large connected component (LCC) that consists of 35,293 nodes (63 % of total nodes) joined by 52,984 edges (76 % of total edges) (Additional file 1: Figure S1). There are 264 disconnected components with more than ten nodes and about 4000 additional small components. Each chromatin fragment has an average of 2.5 neighbors with each promoter interacting with three non-promoter elements on average.

Epigenomic features associated with chromatin fragments participating in 3D contacts

For each fragment in the PCHi-C network, we mapped a large set of 78 epigenomic features [38]. These features included cytosine modifications, histone marks, and ChIP-seq peaks of chromatin-related proteins, such as transcription factors and members of chromatin complexes, including cohesin, CTCF, PcG, and different RNAPII variants (Additional file 2). For each chromatin fragment we calculate the fraction covered by peaks of a specific feature and we define the abundance of each feature as the average of this value over all fragments in the network (see "Methods"). Figure 1a shows the fraction of fragments covered by EZH2 binding sites. We noticed the strong accumulation of the nodes that have binding sites for this PcG factor in specific regions of the network. Strikingly, this co-localization of the signal is observed despite the low overall prevalence of EZH2 binding in the fragments (only 10 % of fragments have some overlap with EZH2 peaks). Figure 1b shows the HoxA cluster region on chromosome 6. In this region, we observe that fragments connected by long-range interactions tend to have similar values of EZH2, with EZH2 peaks having similar heights on pairs of connected fragments. We therefore set out to investigate and quantify the extent to which connected nodes in the whole network have similar values for EZH2 and the other 77 epigenomic features. A high similarity of values in interacting nodes could suggest a role for some features in mediating these contacts.

Definition of ChAs

We propose an approach to identify epigenomic features that can be associated with 3D chromatin contacts. This involves measuring the extent to which neighboring network nodes have similar epigenomic features using ChAs. Assortativity, also called homophily, is the propensity for interacting nodes to have similar values [40] (see "Methods"). ChAs is defined as the assortativity of abundance levels of one specific epigenomic feature on the chromatin interaction network. In practical terms, it is the correlation of abundance of a single feature measured across all pairs of neighbors in the network. As a correlation coefficient, ChAs values range between −1 and 1. ChAs can therefore be used to identify features that are found in fragments that are globally connected in the network or to distinguish different types of fragments that tend to interact with each other. To aid the interpretation of these values, we can consider the three scenarios depicted in a schematic scatter plot of ChAs versus abundance (Fig. 1c):

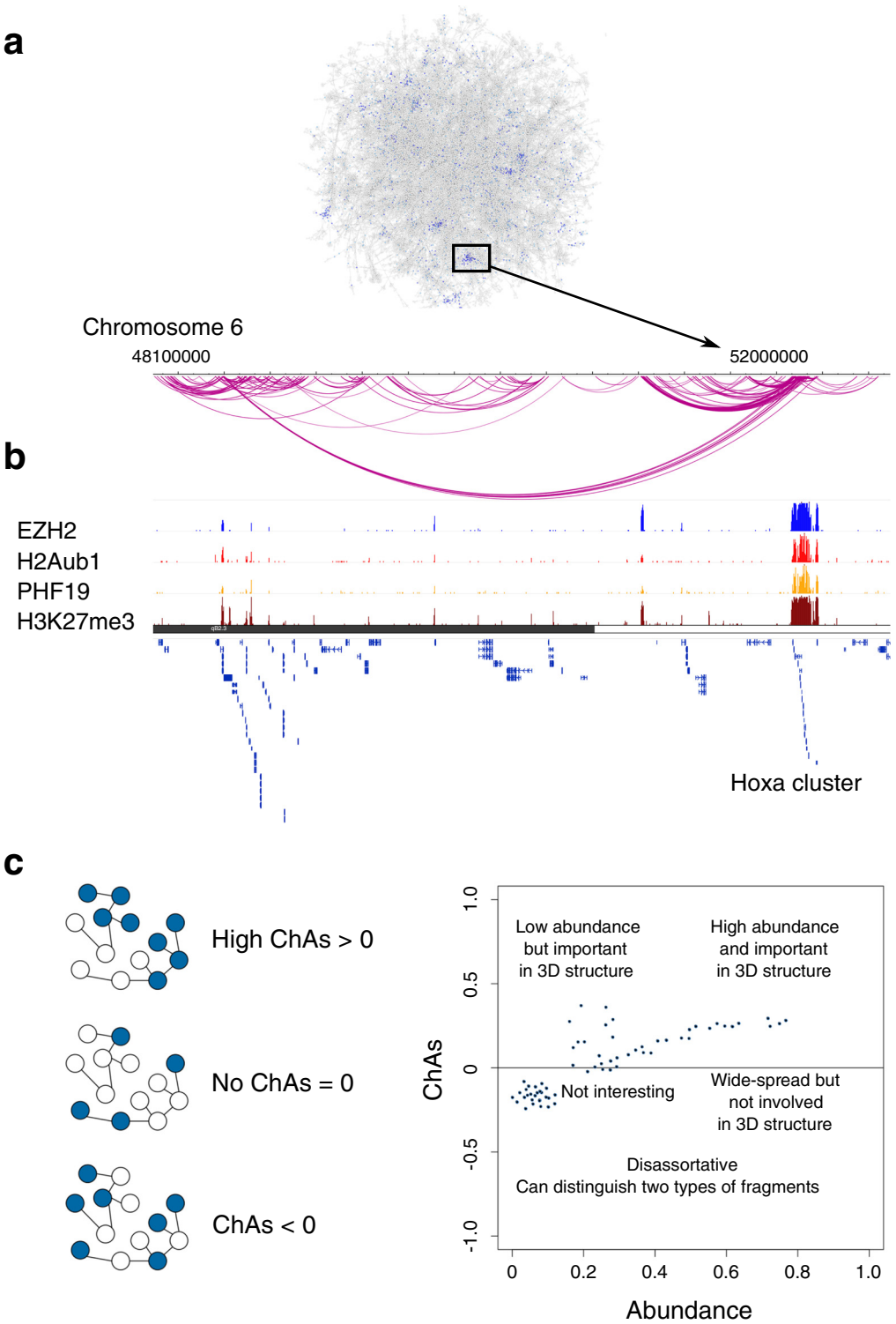

Fig. 1 Chromatin assortativity (ChAs) of epigenomic features in a network of chromatin contacts. **a** Largest connected component of PCHi-C chromatin interaction network in mESCs. Nodes are colored by proportion of fragment covered by EZH2, which highlights the neighborhoods in which the protein is abundant. **b** The genomic region highlighted in the box in (**a**) visualized using the WashU Epigenome browser [67] with added custom tracks for PCHi-C interactions and EZH2 peaks together with other PcG-related features. **c** Cartoon illustrating what ChAs measures. Each *node* of the network is a chromatin fragment, *blue nodes* represent nodes in which a peak of a specific chromatin mark is found, and *edges* represent significant 3D interactions. Next to it we show a cartoon plot of ChAs versus abundance

1. Fragments that have a certain value for the epigenomic feature (that is, certain proportion of the fragment is covered by peaks of that feature) predominantly interact with other fragments which have similar values for the same feature, but not with other fragments. In this case the ChAs for that feature will be positive (ChAs > 0). This situation would indicate that this feature is potentially associated with chromatin contacts.
2. Alternatively, there might be no relationship between the values of the feature on fragments and the values on their neighbors. In this case we will have ChAs = 0. This can happen either when the feature values do not have anything to do with the contacts or when the feature values are very homogeneous in the network: either the feature is low on all fragments (as would be the case for a very rare chromatin mark) or high on all fragments (as would be the case for ubiquitous chromatin marks). This produces low variability of abundance across nodes and, hence, the correlation of these values in neighboring nodes measured by ChAs tends to be 0.
3. Finally, it could be that fragments that have high values for a given feature frequently interact with fragments with low values for that same feature. In this case we will have a negative ChAs (ChAs < 0). This suggests that a set of genomic regions with the feature tend to interact in the network mostly with fragments of a different kind.

For this reason, it is important to consider the abundance of a feature (defined above as the fraction of fragment covered by the feature averaged over fragments) together with the ChAs value. In our EZH2 example, the abundance of this feature is 0.027 and its ChAs value is 0.34, which demonstrates how a fairly rare feature can be assortative.

To summarize, firstly we are interested in features that have high positive ChAs, as this signifies that the mark appears to be localized in specific connected areas of the network. These features are thus very probably involved in the chromatin contacts. Secondly, we are looking for features with negative ChAs, which should be typical of one subclass of fragments that frequently interact with a different subclass of fragments. In this case, ChAs can be used to detect features that distinguish multiple chromatin fragment types.

A recent cohesin ChIA-PET dataset [42] allows us to illustrate the characteristics and biological interpretation of ChAs. Dowen et al. [42] reported interactions with pull-downs of the SMC1 cohesin unit in mESCs. We therefore proceeded to measure abundance and ChAs of SMC1 in this dataset, obtaining a fairly high value of abundance (0.27, mean of all features 0.09) and a low value of ChAs (0.09, mean of all features 0.28). This is expected due to the strong enrichment of fragments for presence of this protein (98 % of fragments have an SMC1 peak). This enrichment makes all fragments have similar proportions covered by the SMC1 feature, hence driving down the ChAs value. CTCF, in contrast, shows an almost threefold increase in ChAs (0.29 versus 0.09 of SMC1) and only a 1.2 % increase in abundance (0.33 versus 0.27 of SMC1) compared with SMC1. These results suggest that the subset of cohesin-bound fragments that also have CTCF bound tend to interact preferentially with each other. In summary, using this well understood dataset, we showed that ChAs is a measure that combines the presence of peaks in different interacting fragments and the topology of the chromatin interaction network. ChAs can thus detect differences and biases in the different types of chromatin interaction networks and identify the chromatin features playing important roles in 3D structure in the cases where these are not known a priori.

ChAs of chromatin features in the mESC chromatin interaction network detected by PCHi-C

We calculated ChAs for the 78 chromatin features in the entire PCHi-C network and compared these values with the corresponding abundance (Fig. 2a). The PcG proteins (EZH2, PHF19, RING1B, SUZ12, CBX7) and histone marks associated with them (H3K27me3, H2Aub1) have the highest ChAs values (ranging from 0.2 to 0.35, mean of all features 0.08; Fig. 2a), suggesting that this complex might be involved in establishing the 3D structure of chromatin in mESCs. This confirms and extends results observed for the Hox gene clusters [8, 20, 41]. RNAPII also has high ChAs, especially the variant implicated in transcriptional elongation (ChAs of RNAPII-S2P = 0.23; Fig. 2a). Two features with high abundance that also have high ChAs are H3K4me1, found on regulatory distal regulatory elements, and H3K36me3, marking transcribed gene bodies. On the other hand, H3K4me3, a modification associated with active promoters, is a very abundant mark (fourth most abundant, abundance = 0.12, mean of all features 0.02) but it has low ChAs (0.04).

We verified that ChAs is robust to random removal of edges in the network, such that our results do not depend on the completeness and accuracy of the chromatin interaction network (see Additional file 1: Text S1 and Figure S2). Moreover, we have ensured the significance of ChAs for at least 72 % of the features using a randomization that preserves network topology and overall feature abundance, as well as using an alternative approach preserving the features' spatial distribution (see Additional file 1: Text S1 and Figure S3). We have also verified that ChAs values are generally not affected by removing short-range contacts that might produce similarity of abundance values in neighboring fragments (Additional file 1: Figures S4 and S5). Finally, comparison of ChAs with other

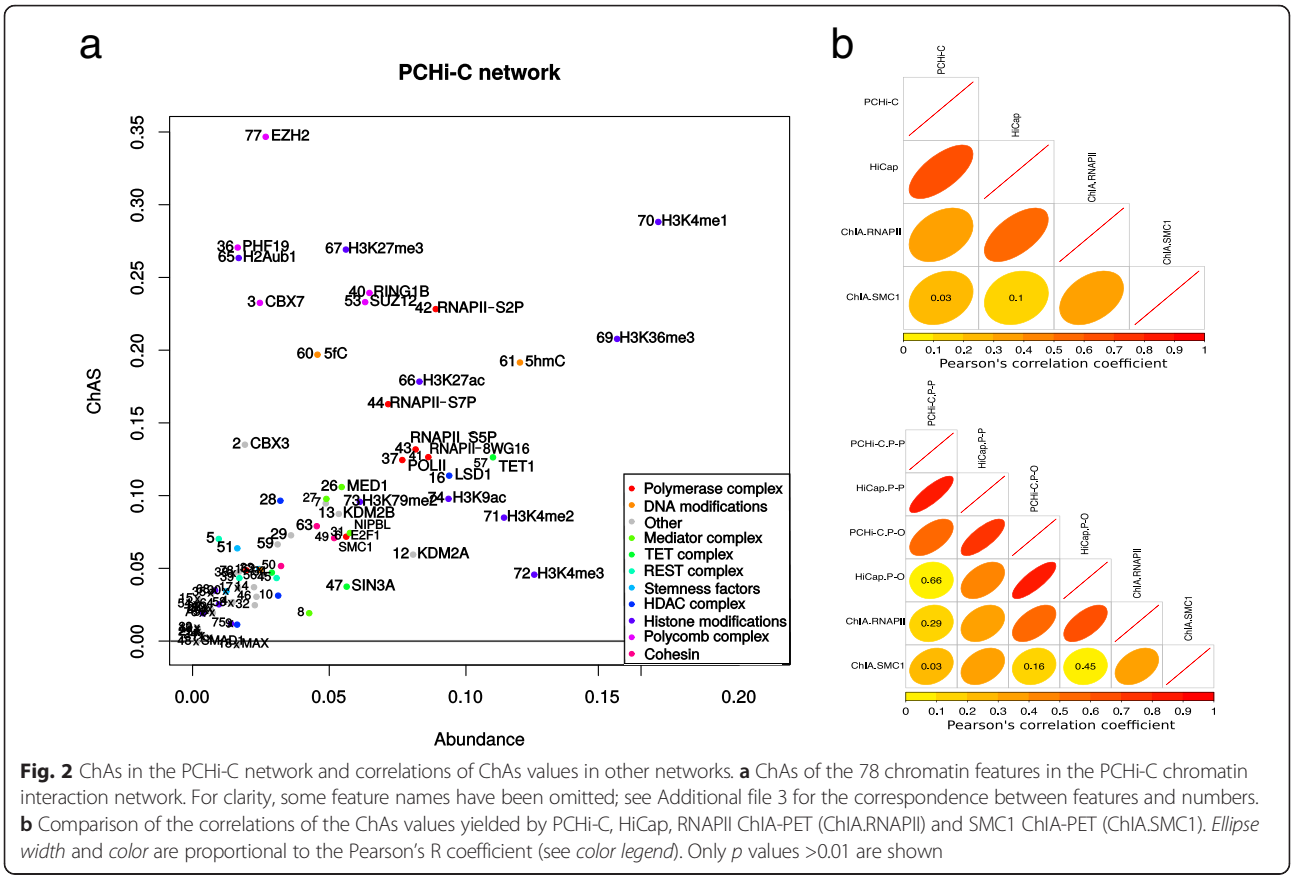

Fig. 2 ChAs in the PCHi-C network and correlations of ChAs values in other networks. **a** ChAs of the 78 chromatin features in the PCHi-C chromatin interaction network. For clarity, some feature names have been omitted; see Additional file 3 for the correspondence between features and numbers. **b** Comparison of the correlations of the ChAs values yielded by PCHi-C, HiCap, RNAPII ChIA-PET (ChIA.RNAPII) and SMC1 ChIA-PET (ChIA.SMC1). *Ellipse width* and *color* are proportional to the Pearson's R coefficient (see *color legend*). Only *p* values >0.01 are shown

network measures demonstrates that it is a complementary method to identify important features (see Additional file 1: Text S2 and S3, Figures S6 and S7).

In summary, the ChAs of an epigenomic feature is a useful global measure that relates feature abundance at interacting fragments with the underlying contact network topology. In the next section, we compare the ChAs values calculated on different chromatin interaction networks.

Chromatin assortativity in additional PCHi-C and ChIA-PET datasets

To test to what extent chromatin interaction network properties depend on the experimental protocol and signal detection algorithm, we took advantage of an alternative promoter interaction dataset in mESCs. Sahlén et al. [7] applied HiCap (a promoter capture method similar to PCHi-C) to mESCs, identifying interactions involving promoters. Using contacts amongst promoters and between promoter and non-promoter fragments from the Sahlén et al. dataset yields a network of 87,823 nodes with 173,801 interactions (including 19,309 promoter nodes and 82,659 P–P interactions). The HiCap technique is complementary to PCHi-C since a different enzyme is used for the restriction step, generating shorter interaction fragments compared with PCHi-C (median size 599 bp versus 3953 bp for PCHi-C). The shorter fragments produce a higher resolution picture of contacts between nearby fragments, at the expense of reduced coverage of long-range interactions. Visualizing the network shows that the largest connected component is comparatively smaller than in PCHi-C, encompassing 9.6 % of the total nodes and 12.8 % of the total connections (Additional file 1: Figure S8).

We analyzed the HiCap network in combination with the 78 chromatin features previously introduced. We repeated the calculation of ChAs of the chromatin features using the HiCap network as described above for the PCHi-C network. We directly compared the ChAs values for all features between PCHi-C and HiCap networks and found that, overall, they are highly correlated (Pearson's R = 0.67, p value = 2.99×10^{-11}; Fig. 2b; Additional file 1: Figure S8d, e). For example, the PcG components are confirmed amongst the features with the highest ChAs, as was observed in the PCHi-C analysis, together with RNAPII, especially the S2P variant (Additional file 1: Figure S8e).

In summary, we have shown that ChAs is a useful metric to detect those epigenomic features that might be more influential in promoter-centered chromatin interaction networks and that the ChAs measurements are rather

independent of the underlying experimental protocol. A comparison with a contact map that is not enriched for contacts involving promoters was performed using the previously mentioned SMC1 ChIA-PET dataset [42] (Additional file 1: Figure S9a, c). There was no significant correlation between ChAs values obtained for the ChIA-PET dataset and the promoter capture datasets (Fig. 2b), showing that the ChAs measurements are specific to the types of contacts assayed (Additional file 1: Figures S10 and S11). The cohesin ChIA-PET network is not enriched for promoters—only 20 % of the SMC1 ChIA-PET fragments overlap the PCHi-C promoter fragments (Additional file 1: Table S1)—but it still shows the assortativity of PcG features and of the actively elongating RNAPII-S2P.

To exclude the possibility that the correlation observed in the two promoter capture datasets was purely due to the experimental technique used to map the contacts, we also calculated ChAs for an RNAPII ChIA-PET dataset. Interactions involving RNAPII (8WG16 antibody, recognizing all variants) have been detected in mESCs [43], allowing us to analyze an RNAPII-focused chromatin interaction network (Additional file 1: Figure S9b, d). In addition, this network allowed us to further test the differences in ChAs of RNAPII variants, which we observed to be reproduced in the PCHi-C, HiCap, and RNAPII ChIA-PET networks but not in the SMC1 ChIA-PET network (Additional file 1: Figures S9–S11). The RNAPII ChIA-PET network is obviously enriched in promoter interactions (58 % of the RNAPII ChIA-PET fragments overlap PCHi-C promoter fragments; Additional file 1: Table S1) but, contrary to the PCHi-C and HiCap promoter-capture networks, it contains only fragments in which RNAPII is bound. Similarly to what we found in the PCHi-C and HiCap networks, PcG proteins and associated histone marks show considerably high ChAs in the RNAPII ChIA-PET network, but lower than H3K4me1 (an enhancer specific mark) and the repressive mark H4K20me3 (Additional file 1: Figure S9b).

The ChAs of the non-specific RNAPII-8WG16 is quite low (0.07) in the RNAPII ChIA-PET network compared with all other features (mean 0.1) (Additional file 1: Figure S9b). A low ChAs is expected given that fragments in this network are highly enriched for the presence of this feature (84 % of fragments have an RNAPII-8WG16 peak, abundance = 0.5). This leads to uniform levels of RNAPII abundance on the nodes and, hence, we do not observe any localization of the mark in specific areas of the contact network. Interestingly, we do observe higher ChAs for the elongating variant RNAPII-S2P (0.19 versus 0.07 for the RNAPII-8WG16) accompanied by a comparatively lower abundance (0.25 versus 0.5 for RNAPII-8WG16), suggesting that regions of the genome in which elongation takes place interact preferentially (Additional file 1: Figure S9b).

Overall, we observe a significant correlation of the RNAPII ChIA-PET ChAs values with PCHi-C (Pearson's R = 0.37, p value = 1.01×10^{-3}; Fig. 2b; Additional file 1: Figure S10c) and an even better correlation with HiCap (Pearson's R = 0.59, p value = 9.77×10^{-9}; Fig. 2b; Additional file 1: Figure S11b), despite the drastically different topology (Additional file 1: Figure S11d).

Comparing the results of our approach using these four different networks, we conclude that the methodology is able to identify the putative roles of specific epigenomic features in mediating different types of chromatin contacts. The high ChAs values of PcG and RNAPII are confirmed in different datasets but different features acquire different levels of ChAs and, potentially, different relevance in the different contact maps. Although PCHi-C, HiCap, and RNAPII ChIA-PET are all enriching for interactions involving promoters, there are clear differences in the resulting networks. Notwithstanding the strong differences in topology and network statistics between promoter-capture and ChIA-PET networks (Additional file 1: Figure S9c–e), we find higher similarity between the three promoter-enriched datasets (PCHi-C, HiCap, and RNAPII ChIA-PET; Additional file 1: Figures S10 and S11). The correlation between ChAs of promoter-capture networks is improved when looking at PCHi-C and HiCap subnetworks that only include P–P contacts or P–O contacts (Fig. 2b; Additional file 1: Figure S12). We therefore proceed with our goal to use ChAs to analyze the difference between interactions involving two promoters and interactions between promoters and other genomic elements.

Distinct ChAs properties of contacts amongst promoters and between promoters and other elements

As mentioned above, the experimental design of promoter-capture HiC (PCHi-C or HiCap) produces chromatin fragments of two kinds: promoter (P) fragments are the ones that are captured in the experiment because they match a library of promoters and are therefore identified as baits; other-end (O) fragments are chromatin fragments found to interact with the promoter baits.

We first investigated the differences in chromatin features associated with PCHi-C contacts involving two promoters (P–P) and contacts involving a promoter and an other-end fragment element (P–O). We calculated feature abundance and ChAs values for two subnetworks: the P–P network and the P–O network (Fig. 3a; Additional file 1: Figure S12). We combined these data in a comparative ChAs plot to directly assess the relationship

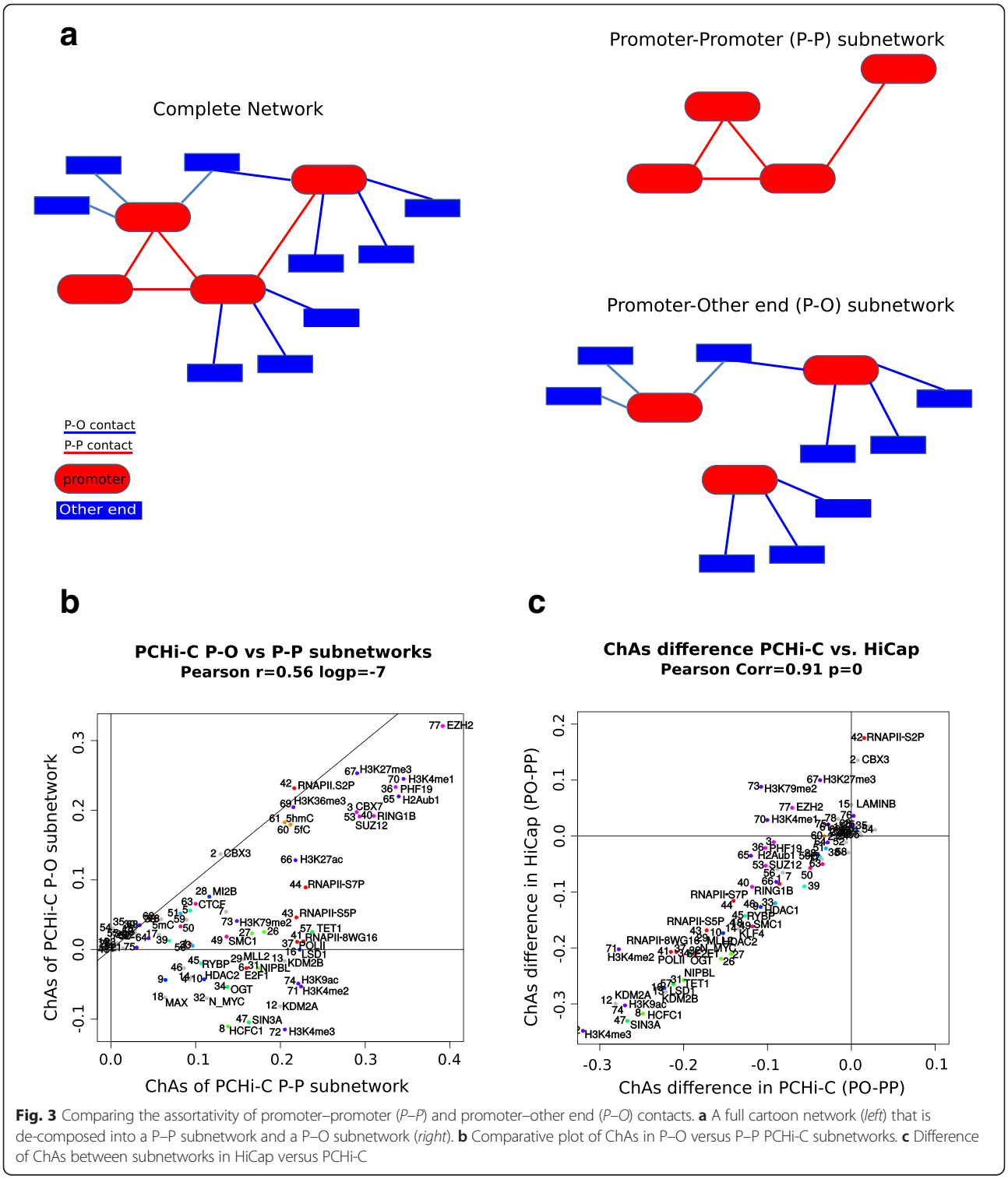

Fig. 3 Comparing the assortativity of promoter–promoter (*P–P*) and promoter–other end (*P–O*) contacts. **a** A full cartoon network (*left*) that is de-composed into a P–P subnetwork and a P–O subnetwork (*right*). **b** Comparative plot of ChAs in P–O versus P–P PCHi-C subnetworks. **c** Difference of ChAs between subnetworks in HiCap versus PCHi-C

between the ChAs of chromatin features measured in the two different subnetworks in PCHi-C (Fig. 3b).

Strikingly, we find a number of features with very different values of ChAs in these two subnetworks. For example, in Fig. 3b we see a group of features with positive ChAs in the P–P interactions, implying that these epigenomic features are found in promoters that contact each other, and negative ChAs in the P–O interaction network, implying that these features are usually not present on the other-end fragments that contact promoters. The features that have discordant signs of ChAs in the two subnetworks include many

promoter-specific histone modifications and chromatin factors, specifically H3K4me3 (typically denoting active promoters), HCFC1 (transcription activator complex), SIN3A (transcriptional repressor complex), KDM2A (H3K26 demethylase), NMYC, OGT (histone acetyl transferase complex), H3K4me2, and H3K9ac (denoting active promoters) [38]. Features that have slightly higher or equal ChAs in the P–O interactions include CBX3 (the HPγ implicated in elongation [44, 45]) and RNAPII-S2P. PCHi-C can only detect interactions involving at least one promoter. At the same time, most of the epigenetic features considered here are characteristic of promoters, due to the large bias in datasets available in the literature. Therefore, we are unlikely to find features with higher ChAs in P–O versus P–P contacts, which would lie at the upper left corner above the diagonal in Fig. 3b. However, the features closer to the diagonal are features that are present in both P–P and P–O contacts. The PcG proteins and their associated histone marks are found very close to the diagonal on the comparative ChAs plot of Fig. 3b, suggesting that they are found at both P–P and P–O contacts, together with H3K36me3 and the cytosine modifications 5hmC and 5fC.

The comparative ChAs plots for the HiCap datasets are very consistent with the PCHi-C ones (Fig. 3c; Additional file 1: Figure S12), as shown clearly in a scatter plot of the difference of ChAs between P–O and P–P subnetworks in the two datasets (Fig. 3c; further comparisons of P–P and P–O ChAs are shown in Additional file 1: Figure S12). Interestingly, we observe substantially different ChAs scores for different RNA polymerase variants exclusively in P–O contacts, with elongating RNAPII having a ChAs 23-fold higher than the non-elongating forms (ChAs of RNAPII-S2P = 0.23 versus 0.01 for RNAPII-8WG16; Fig. 3b).

In order to assess the robustness of these differences, we generated 100 networks by random partial rewiring of the original network and re-calculated the ChAs in P–P and P–O subnetworks (see "Methods" and Additional file 1: Figure S12H). The simulations show non-overlapping simulated ChAs distributions in the P–O subnetworks for the different RNAPII variants, whereas the corresponding distributions in the P–P subnetworks are highly overlapping. These results suggest a significant difference in the role of elongating polymerase between P–P and P–O contacts.

Characterization of overlapping chromatin communities reveals PcG and RNAPII-S2P modules

A large portion of the PCHi-C interactions form a large connected component (LCC), also called a "giant component" [35]. There is a significant correlation of the ChAs values measured for the LCC and for the interactions in the rest of the network (Pearson's r = 0.8, $p = 0$; Additional file 1: Figure S13). However, we observe a higher ChAs for PcG features in the LCC (mean 2.8-fold increase; especially for EZH2, having ChAs = 0.37 in the LCC compared with ChAs = 0.14 in the rest of the network). Considering the LCC, we then identify features that are most abundant in nodes with high betweenness centrality, defined as the number of shortest paths from all nodes to all others that pass through that node [46]. PcG features are enriched in nodes with high betweenness centrality, again suggesting PcG's role in holding the core of the interaction network together (Additional file 1: Figure S14a).

To investigate whether PcG features were also involved in mediating connections between different chromatin communities, or neighborhoods [35], we analyzed the LCC with the ModuLand algorithm, which identifies overlapping modules [47] (Fig. 4a; Additional file 1: Text S3). Once overlapping communities were defined, we calculated the "bridgeness" of each node, defined as the number of different chromatin communities (modules) that it belongs to [48]. Figure 4b shows that the features most abundant in the nodes with highest bridgeness are the ones typical of promoters (SIN3A, HCFC1, and H3K4me3) as well as transcription factors such as E2F1, N-MYC, C-MYC, and KLF4. In contrast, PcG features are not abundant in high bridgeness nodes, suggesting that nodes in which PcG is present do not tend to belong to multiple chromatin communities.

The relative values of bridgeness and betweenness centrality can be used to distinguish the so-called *date* and *party* hubs, defined as nodes that entertain multiple interactions respectively one at a time or simultaneously [49, 50] (Additional file 1: Text S4). Extending this concept and using the enrichment of features in the top bridgeness and betweenness nodes, we can identify "party features", found in nodes that belong to multiple communities at the same time, and "date features", found in nodes involved mainly in one community at any one time (Additional file 1: Figure S14b). Only the PcG features (and to a lesser extent KDM2B, TAF1, and H4K20me3) appear to have a definite "party" character, suggesting that they might mediate more stable interactions due to their high abundance in nodes that are central in the network (high betweenness) but mostly belong to a single community (low bridgeness) (Additional file 1: Figure S14b). Similarly to what was observed for values of ChAs in the P–O subnetwork (Fig. 3b), we see a striking difference between the elongating RNAPII variant S2P and non-elongating RNAPII variants (Fig. 4b; Additional file 1: Figure S14b). The non-elongating RNAPII variants show similarly high abundance in top bridgeness and top betweenness

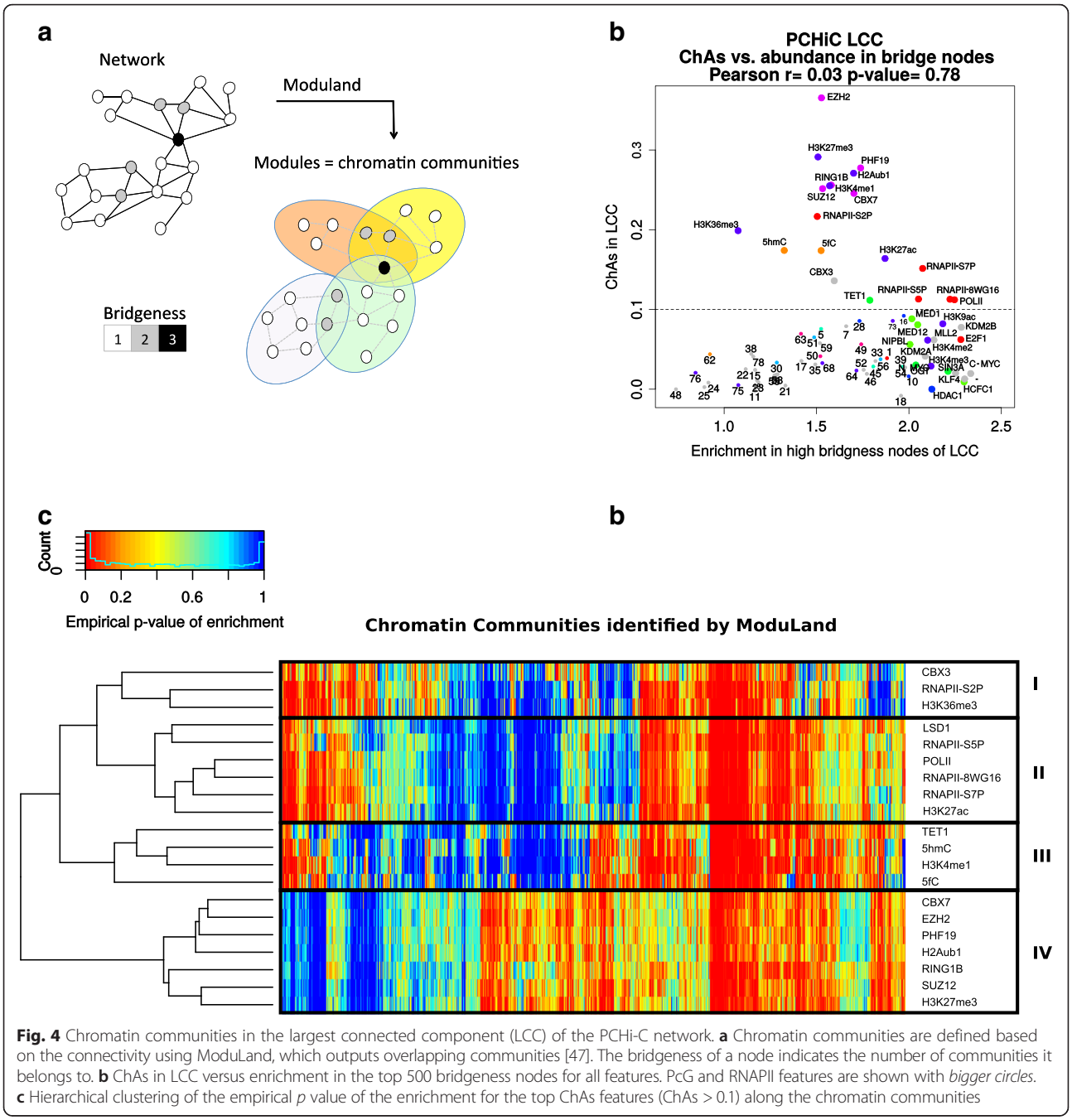

Fig. 4 Chromatin communities in the largest connected component (LCC) of the PCHi-C network. **a** Chromatin communities are defined based on the connectivity using ModuLand, which outputs overlapping communities [47]. The bridgeness of a node indicates the number of communities it belongs to. **b** ChAs in LCC versus enrichment in the top 500 bridgeness nodes for all features. PcG and RNAPII features are shown with *bigger circles*. **c** Hierarchical clustering of the empirical *p* value of the enrichment for the top ChAs features (ChAs > 0.1) along the chromatin communities

nodes, suggesting their presence in nodes that are central and shared between multiple modules. In contrast, the elongating S2P variant is found in more peripheral nodes that specifically belong to a single module, as shown by equally low enrichment in top bridgeness and top betweenness nodes (Additional file 1: Figure S14b). To summarize, PcG features are found in highly connected and highly central nodes, but these nodes do not tend to belong to distinct network communities. The elongating variant of RNAPII, contrary to other RNAPII variants, is found mostly in nodes that belong to a single community and they are more peripheral to the network (low betweenness centrality).

We investigate the difference between RNAPII variants further by looking at enrichment of features in chromatin communities identified by ModuLand, concentrating on the features that showed a high value of ChAs (ChAs > 0.1; Fig. 4b). The heat map in Fig. 4c clearly shows the presence of four clusters. The largest and most prominent is cluster IV including all PcG features, which are enriched in a specific subset of chromatin communities. Clusters II and III contain, respectively,

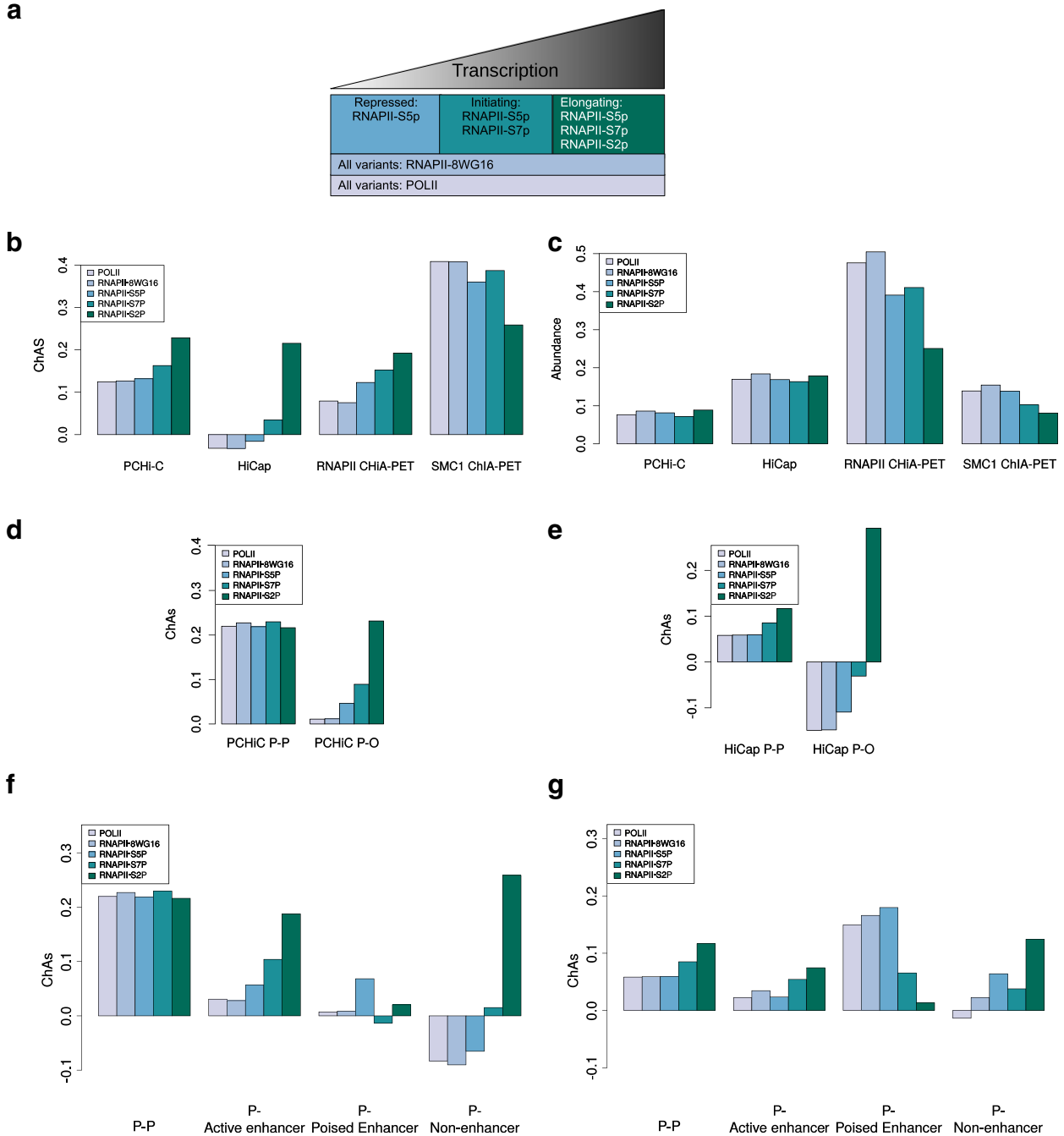

Fig. 5 ChAs of different RNAPII variants in promoter-capture and ChIA-PET networks. **a** Different variants of RNAPII in our chromatin feature set. **b** Comparison of ChAs of RNAPII in PCHi-C, HiCap, RNAPII, and SMC1 ChIA-PET subnetworks. **c** Comparison of abundance of RNAPII in PCHi-C, HiCap, RNAPII, and SMC1 ChIA-PET subnetworks. **d** PCHi-C ChAs in P–P and P–O subnetworks. **e** HiCap ChAs in P–P and P–O subnetworks. **f** PCHi-C ChAs compared between P–P and different types of P–O subnetworks. **g** HiCap ChAs compared between P–P and different types of P–O subnetworks

non-elongating forms of RNAPII and DNA cytosine modifications. On the other hand, RNAPII-S2P appears in cluster I in chromatin communities that are also enriched in H3K36me3 and CBX3. Although all enrichments in RNAPII are anti-correlated with enrichments in PcG features (Fig. 4c), this anti-correlation pattern is stronger for the actively elongating variant RNAPII-S2P (Additional file 1: Figure S15). Overall, these results suggest that PcG features are found in very central and connected nodes that interact stably, forming specific

chromatin communities. Similarly, active elongation is taking place in specific chromatin communities but fragments of chromatin bound by elongating RNAPII are not particularly connected or central in the network (Additional file 1: Figure S6). In the next section we explore the differences between the different RNAPII variants in more detail.

RNAPII-S2P has higher ChAs in P–O contacts compared with other RNAPII variants

Our collection of genome-wide features includes five different ChIP-seq datasets for RNAPII obtained using different antibodies. Of these, three recognize different phosphorylated forms of RNAPII involved in the different stages of transcription [51, 52] (Fig. 5a). We can therefore distinguish between ChIP-seq peaks of RNAPII in its initiating or repressed form (S5P, S7P), in its actively elongating variant (S2P), or in any of its variants (RNAPII-8WG16, POLII).

We compared the ChAs of the different RNAPII variants in the whole PCHi-C and HiCap networks. As was already noted, RNAPII-S2P, which denotes elongation of actively transcribed genes, shows higher ChAs than the other RNAPII variants in both datasets (Fig. 5b). These differences are robust to partial rewiring of the networks (see "Methods" and Additional file 1: Figure S16a). Figure 5c shows the corresponding abundance values, which are very comparable between different RNAPII variants within each dataset.

Next, we compared the ChAs of the different RNAPII variants in the RNAPII ChIA-PET network (Fig. 5b). In principle, the RNAPII ChIA-PET dataset provides us with the network of chromatin contacts in mESCs mediated by any RNAPII, as the antibody used in this experiment (8WG16) recognizes all RNAPII variants. Interestingly, there is an increase of ChAs from repressed to actively elongating RNAPII in all three networks (Fig. 5b; Additional file 1: Figure S16a). These results suggest that, whereas all interacting fragments in these promoter-rich networks do contain some form of polymerase, the presence of active forms of RNAPII distinguishes different network neighborhoods in which active elongation is taking place, as also suggested in Fig. 4c.

Finally, we used the ChIA-PET network of contacts mediated by cohesin in mESCs as a negative control [42]. In this dataset we see many contacts that do not involve any promoters or genes, in which we do not expect to find any RNAPII bound (61 % of fragments in the SMC1 ChIA-PET dataset have no signal for RNAPII-8WG16). Indeed, the different variants of RNAPII in this cohesin-mediated network have very high ChAs (Fig. 5b; Additional file 1: Figure S16a). The presence of any form of RNAPII clearly separates regions of the cohesin-centered network where transcription is active from regions where it is not. These trends cannot be explained by changes in abundance (Fig. 5c).

We further compared the ChAs of different RNAPII variants between P–P and P–O contacts (Fig. 5d). In the PCHi-C network we observe the ChAs for different phosphorylation states of RNAPII to vary widely in the P–O contacts (from close to 0.01 to 0.23, the third highest value overall), while all states have similar ChAs in the P–P contacts (ChAs range 0.21–0.22) (Fig. 5d; Additional file 1: Figure S16b). To understand this trend better, we also look at abundance of the different RNAPII variants in the different subnetworks (Additional file 1: Figure S16c). Whereas in the P–P subnetwork the abundance decreases from inactive forms of RNAPII to the elongating form, in the P–O subnetwork the elongating form is equally abundant compared with the other forms. We can therefore conclude that the different ChAs observed for different forms of RNAPII are related to the topological distribution of RNAPII binding on the network, rather than simply to changes in average abundance in the network. This finding suggests that when O fragments contact P fragments, predominantly the elongating form of RNAPII is present on both fragments. The difference between different RNAPII forms specific to P–O contacts is even more evident in the HiCap dataset where the ChAs value of non-elongating variants of RNAPII is negative (Fig. 5e; Additional file 1: Figure S16d). This is likely due to the higher resolution of the HiCap experiment, which allows us to better discriminate P and O fragments that are probably merged in some of the larger PCHi-C fragments.

We investigated further to determine whether the patterns of ChAs of different RNAPII variants change depending on the type of fragments contacted by the promoter. We selected two types of O fragments: enhancers (fragments with H3K4me1 > 0) divided into active enhancers (H3K4me1 > 0 and H3K27ac > 0) and poised enhancers (H3K4me1 > 0 and H3K27me3 > 0). We can thus separately compare ChAs values between P–P contacts and contacts of P fragments with each type of O fragment. As shown in Fig. 5f, RNAPII-S2P has higher ChAs than the other RNAPII variants in contacts between promoters and active enhancers but not in contacts with poised enhancers (Additional file 1: Figure S16). This suggests that the presence of elongating RNAPII at the P–O contact and the activity of the enhancer might be related.

Strikingly, we also observe a considerable number of contacts between promoters and fragments that do not have the H3K4me1 enhancer mark (H3K4me1 = 0, referred to as non-enhancers in the figure), which we found to be strongly enriched for H3K36me3 (Additional file 1: Figure S17) and that, in 19 % of cases, overlap protein coding gene bodies. In these contacts ChAs

varies from very negative in the non-specific forms to highly positive for the elongating form. This is not due to a change in the abundance of different forms of RNAPII (Additional file 1: Figure S18) and these results are largely confirmed in HiCap (Fig. 5g). These findings suggest that promoters can contact transcribed gene bodies.

Discussion

Assortativity as a robust approach to identify important features in chromatin contacts

We have presented a novel approach, inspired by social network science, which enables the powerful integration of epigenomic features with maps of 3D contacts of chromatin fragments in the nucleus, taking into account the exact network topology. Our approach is robust to the random removal of edges in the contact map, thanks to its global character.

Using the PCHi-C network in mESCs, we demonstrated the capabilities of our assortativity-based approach in recapitulating the importance of PcG factors and associated histone marks. Given the small proportion of fragments that are covered by these marks in the whole genome, the values we observe for their ChAs are highly significant, as also shown by two different randomization procedures. Most features show no change in ChAs value when considering only long-range interactions. PcG features even show higher assortativity in the long-range subnetwork, which is consistent with recent results about PcG mediating extremely long-range contacts [20].

So far, integrated analyses report correlations between genomic information and characteristics of genes in the 3D contact network [4, 10, 53–55], but the exact network topology itself is rarely taken into account. In contrast, the network topology is part of the definition of ChAs and has direct implications in the subsequent calculations. Two very inspiring recent works predict 3D interactions based on 1D epigenomic profiles, but neither provides major insight on the network topology [12, 56].

Having ascertained the appropriateness of chromatin assortativity as a measure, we further define two different subnetworks formed by P–P and P–O interactions and then compare the ChAs for all the features in the two subnetworks. These comparisons show the specific association of certain chromatin features with each type of contact. For example, H3K4me3 has a low ChAs in the complete network but high ChAs in the P–P subnetwork and negative ChAs in the P–O contacts, as corresponds with its role as a differential mark of active promoters.

The *ChAs difference* between the two types of contacts summarises the relationship between features and network topology and permits a direct comparison between datasets. For example, the comparison of ChAs scores between the promoter-capture and the ChIA-PET datasets shows how our method can identify very specific characteristics of the chromatin interaction networks and expose experimental biases. Furthermore, it could be used to identify low quality ChIP-seq datasets, which would fail to show the expected ChAs values.

Biological interpretation of ChAs

We performed this comparison using PCHi-C and HiCap networks to exclude the possibility that our findings are artifacts of one specific dataset. We find a strong correlation of the ChAs values between P–O and P–P subnetworks in the two datasets, giving us confidence in the biological relevance of our results. The reproducibility between the two datasets is remarkable, especially considering the differences in the experimental techniques and the interaction filtering methods used. Whereas PCHi-C is enriched for long-range contacts, HiCap has a higher coverage of short-distance interactions [5, 7, 26], likely constituting connections between promoters and regulatory elements that are relatively close. These types of interactions are probably lost in PCHi-C due to the larger fragment size (which means a single fragment might contain both sites of interaction) and the strict distance correction algorithms applied [29]. Given these differences, the good correspondence of ChAs in the two datasets suggests a general importance for many chromatin factors, which seem to play similar roles in short- and long-range contacts. This is consistent with our observation that ChAs of most features is maintained when removing short-distance contacts (Additional file 1: Figure S5). There are, however, very interesting differences between the ChAs values in P–O contacts in PCHi-C and HiCap, which can be seen by comparing ChAs values directly in the two datasets. For example, more features have negative P–O ChAs values in HiCap. The reason for this is that the larger fragments in PCHi-C will include promoters and also nearby regulatory regions, decreasing the difference between P and O fragment-associated chromatin features.

Looking at the P–P and P–O subnetworks separately also allowed us to notice a marked difference between the variants of RNAPII. The elongating variant appears more strongly associated with contacts between promoters and active enhancers or transcribed gene bodies compared with inactive forms. This is observed in all promoter-centered interaction datasets, including PCHi-C, HiCap, and RNAPII ChIA-PET. In fact, this tendency is given by a decrease in assortativity of the non-elongating RNAPII forms in the contacts between promoters and active enhancers or transcribed gene bodies.

Recently, the presence of RNAPII at distal sites was functionally linked to the activity of CEBP-bound

enhancers, showing that active binding sites display stronger RNAPII binding and local enhancer- RNA production [57]. The presence of polymerase at enhancers was also shown to be strongly predictive for the timing of enhancer activation during development [58]. Our analysis goes beyond these findings and suggests that the presence of non-elongating variants of RNAPII is not associated with preferential contacts of promoters and active distal regulatory elements, whereas the elongating form is. This picture is also consistent with the negative ChAs of non-elongating forms of RNAPII in HiCap P–O contacts. It is possible that the RNAPII that is found at active enhancers is mostly in its elongating form. This is also confirmed by looking at the abundance of RNAPII variants in different fragment types (Additional file 1: Figure S18), which shows that the only form of RNAPII present on other elements is the elongating one. The result is stronger in HiCap contacts, probably because the large size of PCHi-C fragments might signal peaks of RNAPII in O fragments where in reality the peak is in a nearby promoter.

These results are consistent with the different distribution observed between the elongating and non-elongating forms of RNAPII across chromatin communities. Many nodes of the network are found to belong to multiple communities, as evidenced by their high bridgeness. This could indicate that these fragments tend to interact with different partners, either in time or in different cells of the population assayed [31]. The low bridgeness of the elongating form suggests that fragments that are being actively elongated mostly stay within a single chromatin community. Moreover, these fragments are likely to be peripheral to the community itself, given the low betweenness of nodes with high abundance of RNAPII-S2P. This interpretation would be in agreement with the stationary model for RNAPII in transcription factories (assemblies of genes being co-transcribed) [16, 59–61], where elongating RNAPII and nascent transcripts would be localized at the periphery of factories.

We estimated the PcG features to have a more "party hub" than a "date hub" character, given the abundance of these features in the top betweenness and top bridgeness nodes [46, 48]. The concept of date and party hubs is better defined for dynamical networks, typically protein interaction networks in which the former type refers to one-to-one interactions and the latter to stable complexes [48, 49]. In our case we can speculate on the meaning of this distinction, suggesting that PcG features are associated with more stable contacts, which could be more stable both in time and across different cells in a large population [17], and span longer chromosomal distances [20]. In contrast, features present in active promoters and mediating promoter–enhancer contacts are likely to be more specific. The peculiar characteristics of contacts mediated by PcG could be related to the recent observation of major differences between chromatin in the PcG repressed or poised state [62]. These super-resolution microscopy studies found the PcG chromatin to be differently packed from fully active or repressed chromatin, suggesting that the poised domains spatially exclude neighboring active regions.

To summarize these results, we propose the model in Fig. 6, where the network of chromatin contacts (sketched in Fig. 6a) shows regions of promoters that are active, probably due to their contacts with active regulatory elements and transcribed gene bodies. This would lead to high ChAs for the elongating form of RNAPII in both P–P and P–O contacts while ChAs of non-elongating forms would stay low in P–O contacts. Recent literature is suggesting a picture in which enhancer activity is mediated by the formation of loops connecting the gene promoter, the distal enhancer, and the body of the gene [15, 16]. Moreover, 3C experiments have shown that these gene-body contacts are often dynamic and they keep a connection between the gene promoter and the gene body at the exact location of active elongation [63]. We suggest that the RNAPII-S2P variant might be involved in these contacts (Fig. 6b). In the fruit fly, it was proposed that promoter–enhancer contacts are preformed, conserved across tissues and developmental stages, and associated with paused RNAPII [64]. Further experiments will be needed to assess the role of elongation in these processes.

The many interactions we have observed between promoters and gene body fragments without the H3K4me1 enhancer mark cannot be easily explained. It could be speculated that these contacts are joining two promoters while both genes are being transcribed, such that each promoter could come in contact with the body of any of the two genes. This scenario would be consistent with the concomitant localized transcription of multiple genes. This picture is again in line with the concept of transcription factories. Our results on RNAPII-S2P further corroborate this model and are consistent with experimental results showing that whereas the RNAPII-S5P variant would accumulate in the factory, the RNAPII-S2P would remain in the nuclear space nearby the factory [65]. The co-enrichment of modules that we observe for RNAPII-S2P, H3K36me3, and CBX3 (which was shown to interact physically with CDK12 [66], which in turn produces the phosphorylation on RNAPII necessary for active elongation) further support the separation of fragments being actively elongated from the transcription factory.

Conclusions

We have demonstrated the use of assortativity of chromatin features in interpreting chromatin interaction

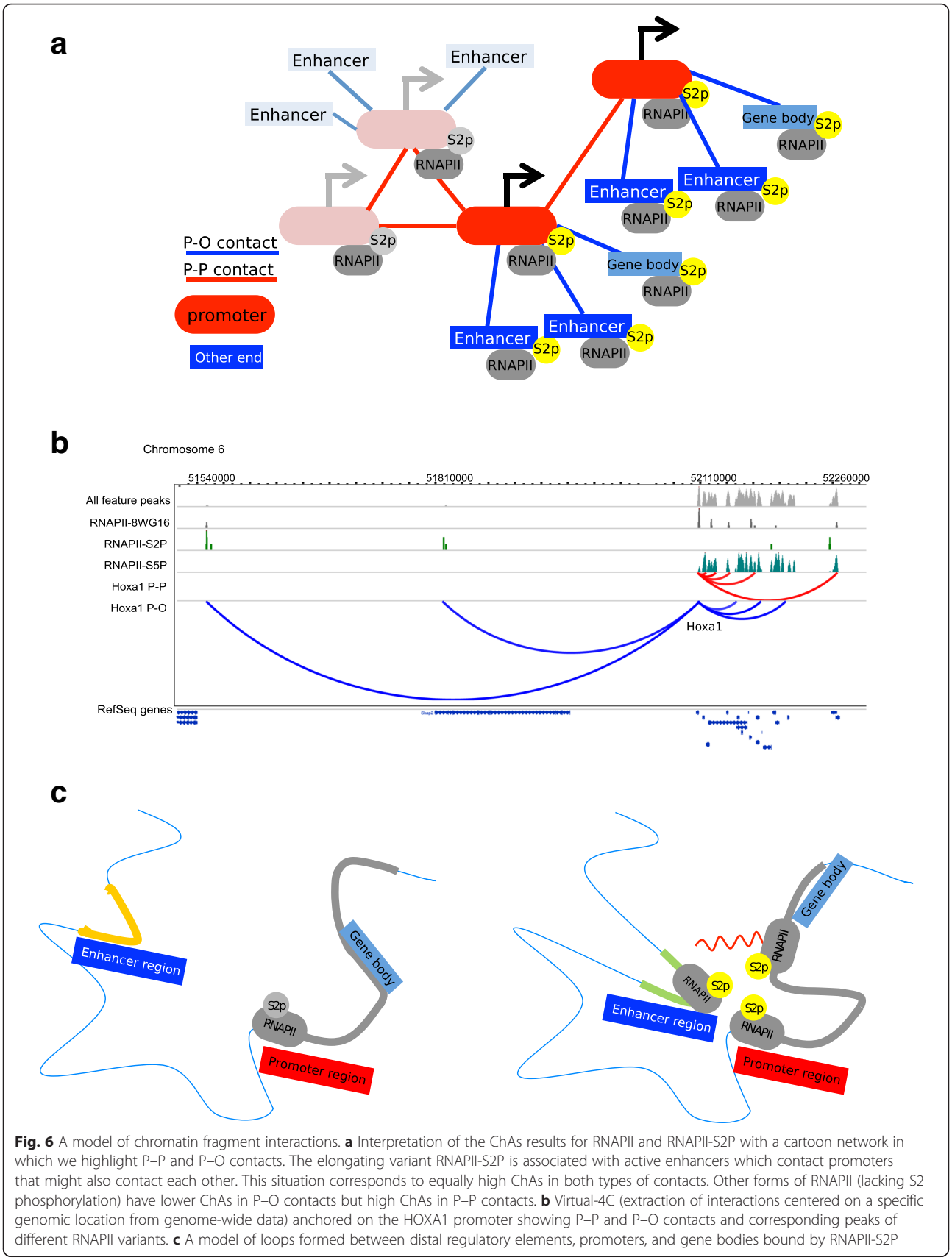

Fig. 6 A model of chromatin fragment interactions. **a** Interpretation of the ChAs results for RNAPII and RNAPII-S2P with a cartoon network in which we highlight P–P and P–O contacts. The elongating variant RNAPII-S2P is associated with active enhancers which contact promoters that might also contact each other. This situation corresponds to equally high ChAs in both types of contacts. Other forms of RNAPII (lacking S2 phosphorylation) have lower ChAs in P–O contacts but high ChAs in P–P contacts. **b** Virtual-4C (extraction of interactions centered on a specific genomic location from genome-wide data) anchored on the HOXA1 promoter showing P–P and P–O contacts and corresponding peaks of different RNAPII variants. **c** A model of loops formed between distal regulatory elements, promoters, and gene bodies bound by RNAPII-S2P

datasets in the context of available epigenomic data. We have achieved this through the definition of ChAs, a measure of how much the value of a specific chromatin feature is correlated between a chromatin fragment and others that interact preferentially with it. The difference of ChAs between the P–P and P–O subnetworks can be used to compare two or more chromatin interaction datasets. Thus, the method we present provides a quick and efficient comparison of different chromatin contact networks and integration of complementary epigenomic and functional information.

Comparing two different networks obtained with two variations of promoter capture HiC on mESCs, we find excellent reproducibility of the following observations: (1) members of the PcG and associated marks show high ChAs despite the low abundance in the interacting fragments, suggesting that they mediate the 3D contacts, especially in the long-range, as already noted [20]; (2) the ChAs values of different variants of RNAPII suggest a picture in which contacts happen between enhancers, their target promoters, and along the gene body. Moreover, we identify the important role of the actively elongating variant of RNAPII in interactions between promoters, distal elements, and other sites in the gene body. Whether it is the contact between these different chromatin regions that spreads the localization of RNAPII-S2P or RNAPII in its elongating form that promotes the contacts remains to be examined in further work.

ChAs is a new complementary measure that provides a global view based on integrating network topology with feature values in the interacting fragments. It has recently been suggested that features located within the loop connecting promoter and enhancer can be determining for the interaction [12], which suggests that expanding our analysis to combine HiC- and PCHi-C-derived networks might yield further insight.

Our results across four different chromatin interaction networks, spanning different techniques and identifying different biases, lend support to the presented ChAs approach as a useful tool in the quest for organizing principles shaping chromatin contact networks.

Methods

Generating the PCHi-C network

PCHi-C interactions measured in mESCs in [8] were processed using CHiCAGO [29]. The publicly available HiCUP pipeline (Wingett et al., submitted) was used to process the raw sequencing reads. This pipeline was used to map the read pairs against the mouse (mm9) genome, to filter experimental artifacts (such as circularized reads and re-ligations), and to remove duplicate reads. The resulting BAM files were processed into CHiCAGO input files, retaining only those read pairs that mapped, at least on one end, to a captured bait.

CHiCAGO is a method to detect significant HiC interactions specifically adapted to promoter capture experiments. In brief, it uses a noise convolution model in which two noise terms account independently for noise sources that dominate at different scales: (1) Brownian motion, which leads to probabilities of interactions decreasing with distance and is modeled using a negative binomial distribution; and (2) sequence artifacts, which are modeled using a Poisson distribution. Once the ChiCAGO scores had been obtained, only interactions with a score ≥ 5 were considered.

The network was then built considering each fragment as a node (therefore having two types of nodes, namely promoters and other ends, and two types of edges, namely promoter–other end and promoter–promoter. Multiple edges connecting the same two nodes were eliminated.

HiCap and ChIA-PET networks

The HiCap data were downloaded from the supplementary material of Sahlén et al. [7], which provides coordinates of the promoter and other end fragments that show significant interaction as well as a list of gene promoters that interact based on assignation of promoter fragments to the closest transcription start site. Interactions not involving promoter fragments were filtered out. The fragment coordinates and interactions of the SMC1 ChIA-PET dataset were downloaded from the supplementary material of [42]. The fragment coordinates of the RNAPII ChIA-PET dataset were downloaded from the supplementary material of [43]. No further processing or filtering was made for these two datasets.

Calculation of feature abundance in the chromatin fragments

The chromatin features (Additional file 2) were taken from Juan et al. [38] and the peak-calling (binarization) was performed as described there in 200-bp windows. For each fragment the overlapping windows of chromatin peaks were identified and their values averaged to give a fraction of presence of any feature in each fragment. Thus, for each feature a value between 0 and 1 is associated with each fragment (which has an average length of 4.9 kb in PCHi-C and 600 bp in HiCap), generating a fragment-by-feature matrix. The value of abundance of a feature is defined as the average of that feature value across all fragments in the network considered.

ChAs calculation

We define the ChAs of a specific epigenomic feature in a contact network as the Pearson correlation coefficient of the value of that feature across all pairs of nodes that are connected with each other [40]. ChAs is, therefore, the assortativity of the abundance value of a feature on a

network. We used the igraph (version 0.7.1) package in R and its function "assortativity" to calculate the ChAs of each feature separately on the network of choice (either the full, the P–P, or the P–O network) in PCHi-C and HiCap.

The assortativity measure used was that for continuous variables given by the following formula:

$$r = \sum_{jk} jk\left(e_{jk} - q_j q_k\right)/\sigma_q^2$$

where $q_i = \sum_j e_{ij}$, e_{ij} is the fraction of edges connecting vertices of type i and j, and σ_q is the standard deviation of q.

A more intuitive definition of assortativity is simply the Pearson correlation between two vectors: vector 1 contains the feature values of the source nodes and vector 2 contains the feature values of the sink nodes, once all edges in the network are enumerated. There is no appreciable difference in the value of assortativity obtained by listing all edges in an arbitrary direction or first adding all edges in the opposite direction and calculating assortativity on this extended network.

Robustness and significance of ChAs values

We assessed how the ChAs values can be affected by the accuracy of the topology of the chromatin interaction network by removing edges at random and following targeted approaches based on feature abundance. Further details and results can be found in Additional file 1: Text S1 and Figure S2.

We also tested whether the ChAs of the chromatin features we measured was significantly higher than would be expected at random using two different approaches. Briefly, in the first approach we shuffled the assignment of feature values between network nodes, repeating this 100 times and thus calculating empirical p values. In the second approach we created new interactions between bait fragments of chromosome 1 and randomly chosen regions of the same chromosome, with the same size and distance from bait as the original other-end fragments, also calculating empirical p values. Further details, a schematic description of the two approaches, and results can be found in Additional file 1: Text S1 and Figures S3 and S4.

Finally, to assess the impact of differences between ChAs of different features in the same network or the same feature across networks, we performed a partial rewiring of the networks and calculated the distribution of ChAs values for each feature (10 % of edges swapped).

Network analyses and community detection

Network properties such as degree, transitivity, betweenness centrality, and number of connected components were calculated using igraph. Further details on the network analyses and results can be found in Additional file 1: Text S2 and Figure S6.

We identified chromatin communities in the PCHi-C network using two separate approaches. First, we used the ModuLand plugin for Cytoscape [47], which returns overlapping network communities and values of bridgeness for each network node (defined as number of communities that the node belongs to [46]). Second, we used a fast greedy community finding algorithm from the igraph package to identify non-overlapping network modules. Further details on the community detection and results can be found in Additional file 1: Text S3 and Figure S7.

Definition of different types of O fragments

Active enhancers were defined as other-end fragments with the value of H3K4me1 > 0 and H3K27ac > 0. Poised enhancers were defined as other-end fragments with the value of H3K4me1 > 0 and H3K27me3 > 0. For example, given our definition of feature abundance, this will identify an active enhancer in any fragment that has at least one 200-bp segment covered by a H3K4me1 peak and at least one (not necessarily the same) 200-bp segment covered by a H3K27ac peak. We have identified non-enhancers as O fragments for which the value of H3K4me1 = 0.

Statistical analyses

All analyses were performed using R version 3.1.0 (x86_64-pc-linux-gnu) (R Development Core Team 2008).

Additional files

Additional file 1: A PDF file containing all supplementary texts, figures and tables. (PDF 4177 kb)

Additional file 2: An .xls file containing all features used and references to the datasets [38]. (XLS 53 kb)

Additional file 3: A tab-formatted text file containing the calculated values of abundance and ChAs for all features in all four networks. (TXT 23 kb)

Additional file 4: A tab-formatted text file containing the results of the PCHi-C randomization preserving network topology. (TXT 157 kb)

Additional file 5: A tab-formatted text file containing the results of the PCHi-C randomization preserving feature distributions along the chromosomes. (TXT 74 kb)

Abbreviations

3D, three-dimensional; ChAs, chromatin assortativity; ChIA-PET, chromatin interaction analysis by paired-end tag sequencing; ChICAGO, Capture Hi-C Analysis of Genome Organization; HiC, high-throughput conformation capture; LCC, large connected component; mESC, mouse embryonic stem cell; O, other-end; P, promoter; PcG, Polycomb group; PCHiC, promoter capture HiC; RNAPII, RNA polymerase II

Acknowledgements

We thank Simone Marsili and two anonymous referees for providing interesting comments and improvements to our work.

Funding
The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 282510 (BLUEPRINT) and in the framework of the Platform for Biomolecular and Bioinformatics Resource PT 13/0001/0030 of the ISCIII, which is funded through the European Regional Development Fund (ERDF). This work was also supported by grants from the Biotechnology and Biological Sciences Research Council BBS/E/B/000C0404 and the Medical Research Council MR/L007150/1 to PF. VP acknowledges a FEBS Long-term fellowship.

Availability of data and materials
The chromatin features used were already compiled from the literature in [38] and can be found in Additional file 2. A table with calculated values of abundance and ChAs for all the networks used and all the features can be found in Additional file 3. Results of the randomization approaches can be found in Additional files 4 and 5 and abundances for fragments in the different networks can be found on the figshare website. The code and networks are available at https://figshare.com/s/c331ded300fefd8f3ec0.

Authors' contributions
VP performed the analyses; EC, VP, and MS processed the different datasets; VP, EC, BMJ, DJ, and DR interpreted the results. VP and DR devised the study and drafted the initial manuscript. DR, MS, PF, and AV supervised the work. All authors revised and approved the final version of the manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [2]Nuclear Dynamics Programme, The Babraham Institute, Cambridge, UK.

Received: 10 May 2016 Accepted: 7 June 2016
Published online: 08 July 2016

References

1. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of chromosomal domains by loop extrusion. bioRxiv. 2015: 024620.
2. Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, Jewett AI, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci U S A. 2015;112: E6456–65.
3. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.
4. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. 2015;47:598–606.
5. Schoenfelder S, Sugar R, Dimond A, Javierre B-M, Armstrong H, Mifsud B, et al. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. Nat Genet. 2015;47:1179–86.
6. Pombo A, Dillon N. Three-dimensional genome architecture: players and mechanisms. Nat Rev Mol Cell Biol. 2015;16:245–57.
7. Sahlén P, Abdullayev I, Ramsköld D, Matskova L, Rilakovic N, Lötstedt B, et al. Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. Genome Biol. 2015;16:156.
8. Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre B-M, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. 2015;25:582–97.
9. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.
10. Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, et al. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol Syst Biol. 2015;11:852.
11. Mirny LA. The fractal globule as a model of chromatin architecture in the cell. Chromosome Res. 2011;19:37–51.
12. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet. 2016;48:488–96.
13. Xu C, Corces VG. Towards a predictive model of chromatin 3D organization. Semin Cell Dev Biol. 2015.
14. Dekker J, Mirny L. The 3D genome as moderator of chromosomal communication. Cell. 2016;164:1110–21.
15. Andersson R. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. Bioessays. 2015;37:314–23.
16. Feuerborn A, Cook PR. Why the activity of a gene depends on its neighbors. Trends Genet. 2015;31:483–90.
17. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013;502:59–64.
18. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, et al. Chromatin architecture reorganization during stem cell differentiation. Nature. 2015;518:331–6.
19. Krijger PHL, Di Stefano B, de Wit E, Limone F, van Oevelen C, de Laat W, et al. Cell-of-origin-specific 3D genome structure acquired during somatic cell reprogramming. Cell Stem Cell. 2016.
20. Joshi O, Wang S-Y, Kuznetsova T, Atlasi Y, Peng T, Fabre PJ, et al. Dynamic reorganization of extremely long-range promoter-promoter interactions between two states of pluripotency. Cell Stem Cell. 2015;17:748–57.
21. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485:376–80.
22. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. 2015;10:1297–309.
23. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Bin MY, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009;462:58–64.
24. Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. Genome Res. 2014;24:1854–68.
25. Jäger R, Migliorini G, Henrion M, Kandaswamy R, Speedy HE, Heindl A, et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nat Commun. 2015;6:6178.
26. Martin P, McGovern A, Orozco G, Duffus K, Yarwood A, Schoenfelder S, et al. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. Nat Commun. 2015;6:10069.
27. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods. 2015;72:65–75.
28. Ay F, Noble WS. Analysis methods for studying the 3D architecture of the genome. Genome Biol. 2015;16:183.
29. Cairns J, Freire-Pritchett P, Wingett SW, Dimond A, Plagnol V, Zerbino D, et al. CHiCAGO: robust detection of DNA looping interactions in capture Hi-C data. Genome Biol. 2016;17:127. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0992-2
30. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16:259.
31. Sekelja M, Paulsen J, Collas P. 4D nucleomes in single cells: what can computational modeling reveal about spatial chromatin conformation? Genome Biol. 2016;17:54.
32. Nicodemi M, Pombo A. Models of chromosome structure. Curr Opin Cell Biol. 2014;28:90–5.
33. Hoang SA, Bekiranov S. The network architecture of the Saccharomyces cerevisiae genome. PLoS One. 2013;8, e81972.
34. Babaei S, Mahfouz A, Hulsman M, Lelieveldt BPF, de Ridder J, Reinders M. Hi-C chromatin interaction networks predict co-expression in the mouse cortex. PLoS Comput Biol. 2015;11, e1004221.
35. Sandhu KS, Li G, Poh HM, Quek YLK, Sia YY, Peh SQ, et al. Large-scale functional organization of long-range chromatin interaction networks. Cell Rep. 2012;2:1207–19.
36. Kruse K, Sewitz S, Babu MM. A complex network framework for unbiased statistical analyses of DNA-DNA contact maps. Nucleic Acids Res. 2012;41:701–10.
37. Singh A, Bagadia M, Sandhu KS. Spatially coordinated replication and minimization of expression noise constrain three-dimensional organization of yeast genome. DNA Res. 2016;23:155–69.

38. Juan D, Perner J, Carrillo de Santa Pau E, Marsili S, Ochoa D, Chung H-R, et al. Epigenomic co-localization and co-evolution reveal a key role for 5hmC as a communication hub in the chromatin network of ESCs. Cell Rep. 2016;14:1246–57.
39. McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: homophily in social networks. Annu Rev Sociol. 2001;27:415–44.
40. Newman MEJ. Assortative mixing in networks. Phys Rev Lett. 2002;89: 208701.
41. Denholtz M, Bonora G, Chronis C, Splinter E, de Laat W, Ernst J, et al. Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. Cell Stem Cell. 2013;13:602–16.
42. Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell. 2014;159:374–87.
43. Zhang Y, Wong C-H, Birnbaum RY, Li G, Favaro R, Ngan CY, et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature. 2013;504:306–10.
44. Smallwood A, Hon GC, Jin F, Henry RE, Espinosa JM, Ren B. CBX3 regulates efficient RNA processing genome-wide. Genome Res. 2012;22:1426–36.
45. Vakoc CR, Mandat SA, Olenchock BA, Blobel GA. Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin. Mol Cell. 2005;19:381–91.
46. Newman MEJ. Networks. An introduction. Oxford: Oxford University Press; 2010.
47. Szalay-Beko M, Palotai R, Szappanos B, Kovács IA, Papp B, Csermely P. ModuLand plug-in for Cytoscape: determination of hierarchical layers of overlapping network modules and community centrality. Bioinformatics. 2012;28:2202–4.
48. Kovács IA, Palotai R, Szalay MS, Csermely P. Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. PLoS One. 2010;5, e12528.
49. Han J-DJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 2004;430:88–93.
50. Kovács IA, Palotai R, Szalay MS, Csermely P. Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. PLoS One. 2010;5:14.
51. Brookes E, de Santiago I, Hebenstreit D, Morris KJ, Carroll T, Xie SQ, et al. Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. Cell Stem Cell. 2012;10:157–70.
52. Jonkers I, Lis JT. Getting up to speed with transcription elongation by RNA polymerase II. Nat Rev Mol Cell Biol. 2015;16:167–77.
53. Akdemir KC, Chin L. HiCPlotter integrates genomic data with interaction matrices. Genome Biol. 2015;16:198.
54. Merelli I, Liò P, Milanesi L. NuChart: an R package to study gene spatial neighbourhoods with multi-omics annotations. PLoS One. 2013;8, e75146.
55. Moore BL, Aitken S, Semple CA. Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization. Genome Biol. 2015;16:110.
56. Zhu Y, Chen Z, Zhang K, Wang M, Medovoy D, Whitaker JW, et al. Constructing 3D interaction maps from 1D epigenomes. Nat Commun. 2016;7:10812.
57. Savic D, Roberts BS, Carleton JB, Partridge EC, White MA, Cohen BA, et al. Promoter-distal RNA polymerase II binding discriminates active from inactive CCAAT/enhancer-binding protein beta binding sites. Genome Res. 2015;25:1791–800.
58. Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, Delhomme N, et al. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat Genet. 2012;44:148–56.
59. Rieder D, Trajanoski Z, McNally JG. Transcription factories. Front Genet. 2012;3:221.
60. Sutherland H, Bickmore WA. Transcription factories: gene expression in unions? Nat Rev Genet. 2009;10:457–66.
61. Chakalova L, Debrand E, Mitchell JA, Osborne CS, Fraser P. Replication and transcription: shaping the landscape of the genome. Nat Rev Genet. 2005;6:669–77.
62. Boettiger AN, Bintu B, Moffitt JR, Wang S, Beliveau BJ, Fudenberg G, et al. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. Nature. 2016;529:418–22.
63. Lee K, Hsiung CC-S, Huang P, Raj A, Blobel GA. Dynamic enhancer-gene body contacts during transcription elongation. Genes Dev. 2015;29:1992–7.
64. Ghavi-Helm Y, Klein FA, Pakozdi T, Ciglar L, Noordermeer D, Huber W, et al. Enhancer loops appear stable during development and are associated with paused polymerase. Nature. 2014;512:96–100.
65. Ghamari A, van de Corput MPC, Thongjuea S, van Cappellen WA, van Ijcken W, van Haren J, et al. In vivo live imaging of RNA polymerase II transcription factories in primary cells. Genes Dev. 2013;27:767–77.
66. Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, et al. Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. Cell. 2006;125:173–86.
67. Zhou X, Lowdon RF, Li D, Lawson HA, Madden PAF, Costello JF, et al. Exploring long-range genome interactions using the WashU Epigenome Browser. Nat Methods. 2013;10:375–6.